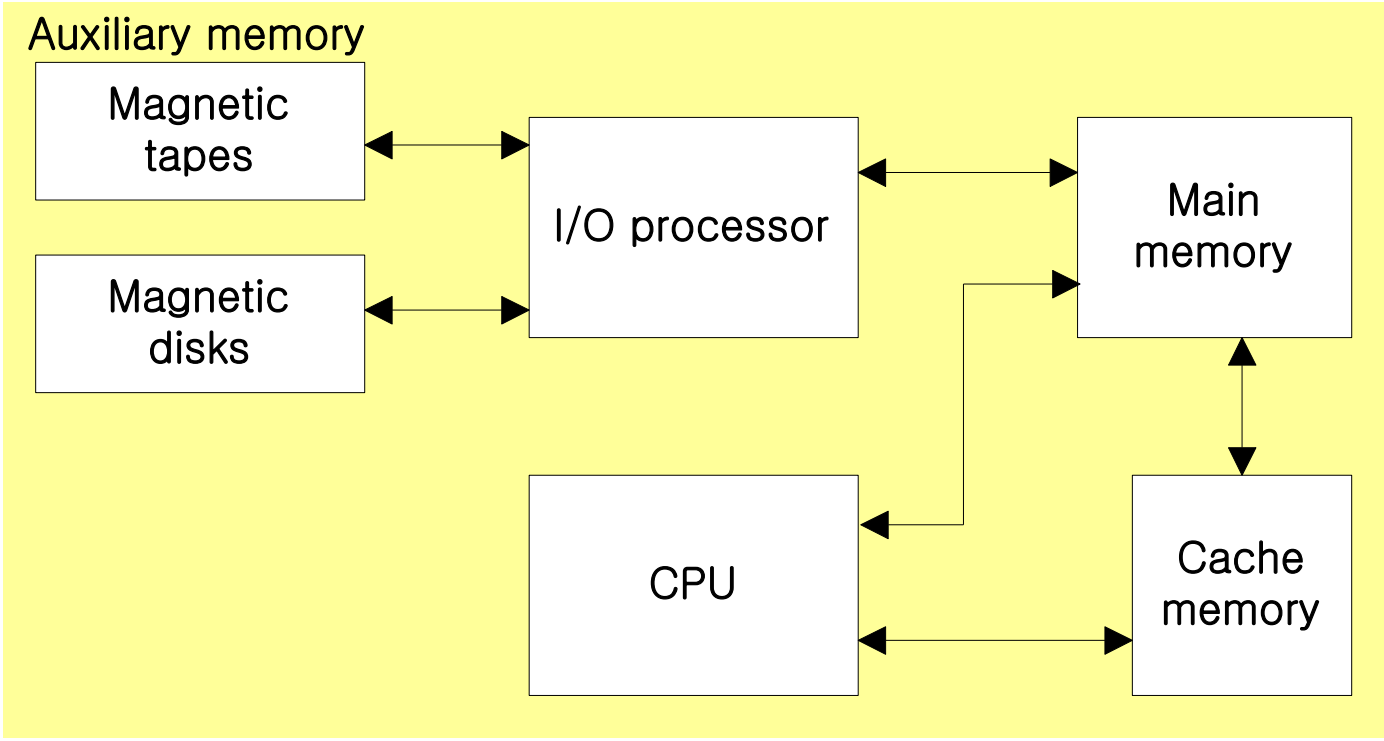
Memory Organization

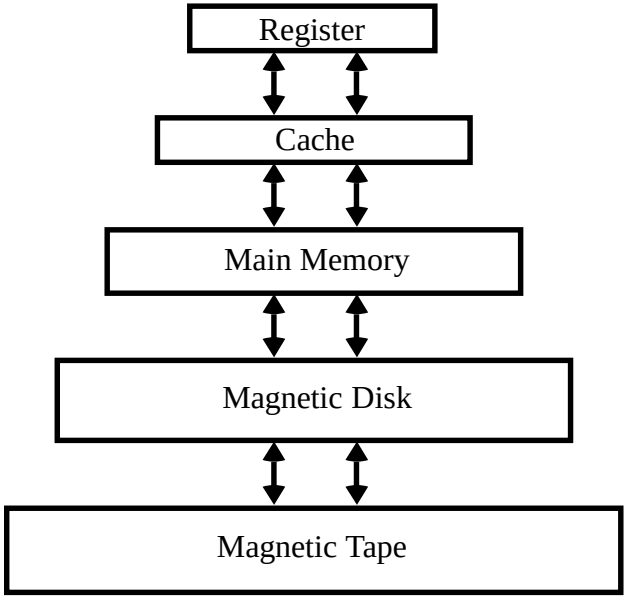
- Memory Hierarchy
  - Memory hierarchy in a computer system :
    - **Main Memory** : memory unit that communicates directly with the CPU (RAM)
    - **Auxiliary Memory** : device that provide backup storage (Disk Drives)
    - **Cache Memory** : special very-high-speed memory to increase the processing speed (Cache RAM)



Memory Organization

- Multiprogramming
  - enable the CPU to process a number of independent program concurrently
- Memory Management System :
  - supervise the flow of information between auxiliary memory and main memory

- Memory Hierarchy is to obtain the highest possible access speed while minimizing the total cost of the memory system



Random Access Memory (RAM)

RAM Chips

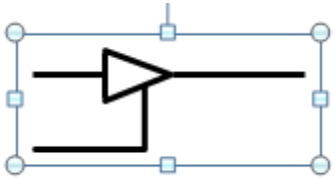
Static RAM

- Consists of Flip-flops to store binary information.
- Static RAM is easier to use and having short Read/Write cycles
- Used mostly in Cache memory.

Dynamic RAM

- Stores binary information in the form of electric charge stored inside the capacitor.
- The capacitors are provided by MOS transistors.
- Refreshing circuit is required to refresh the memory.
- Dynamic RAM offers reduced power consumption and large storage capacity.
- It is used to construct main memory.

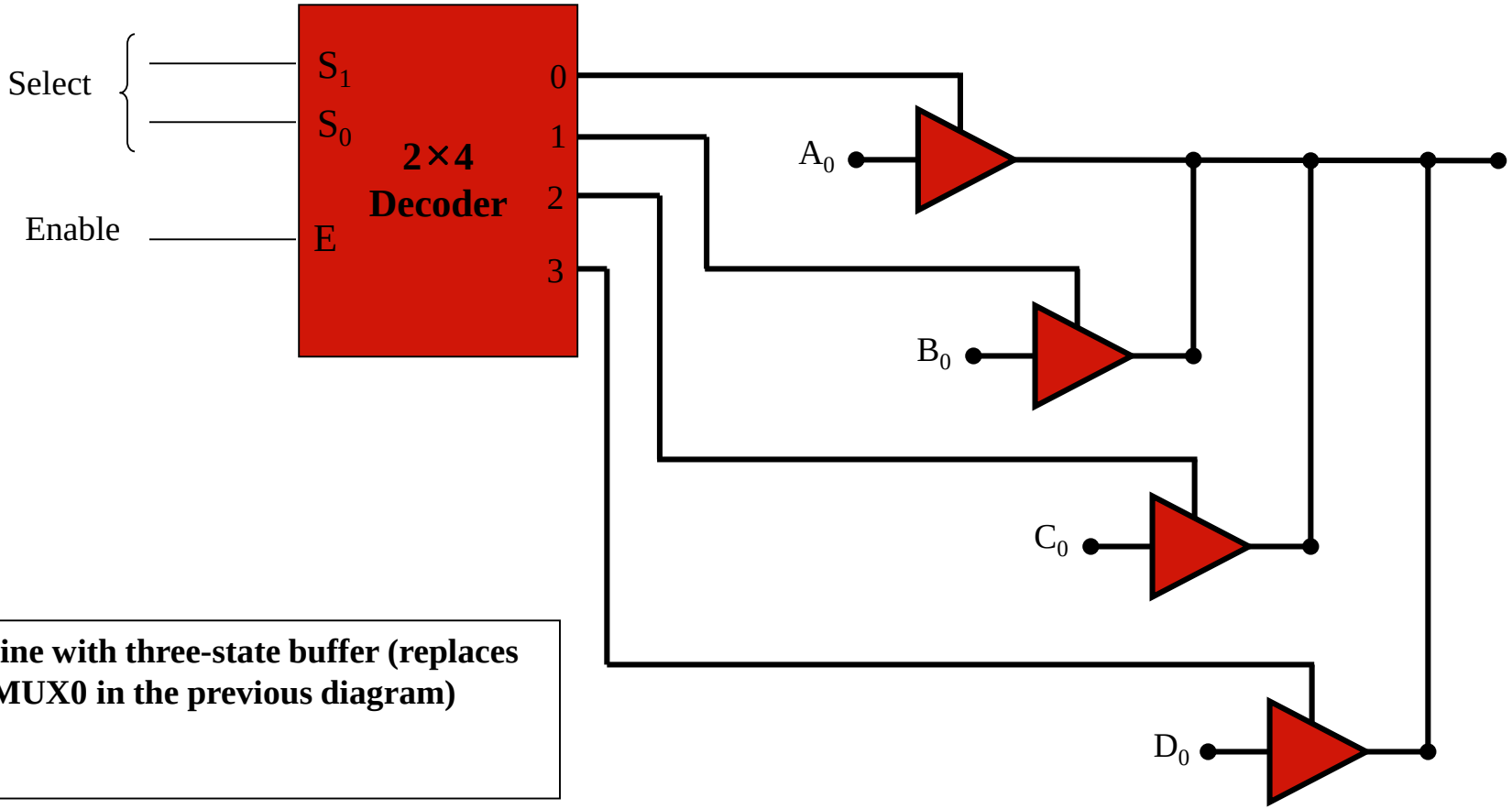
Three State Buffer



Normal input A

Control input C

Output Y=A if C=1
High-impedence if C=0

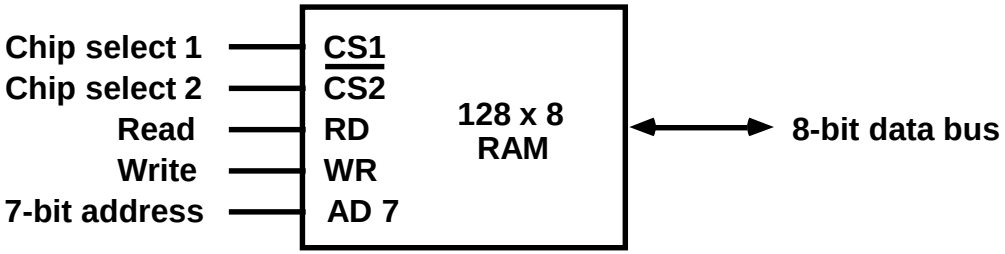


Bus line with three-state buffer (replaces MUX0 in the previous diagram)

MAIN MEMORY

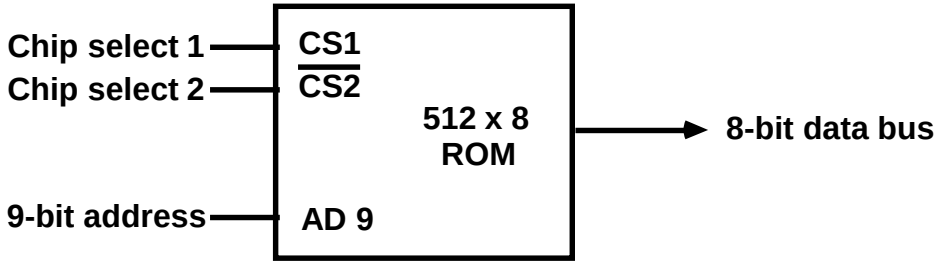
RAM and ROM Chips

Typical RAM chip



CS1	$\overline{CS2}$	RD	WR	Memory function	State of data bus
0	0	x	x	Inhibit	High-impedence
0	1	x	x	Inhibit	High-impedence
1	0	0	0	Inhibit	High-impedence
1	0	0	1	Write	Input data to RAM
1	0	1	x	Read	Output data from RAM
1	1	x	x	Inhibit	High-impedence

Typical ROM chip



MEMORY ADDRESS MAP

Address space assignment to each memory chip

Example: 512 bytes RAM and 512 bytes ROM

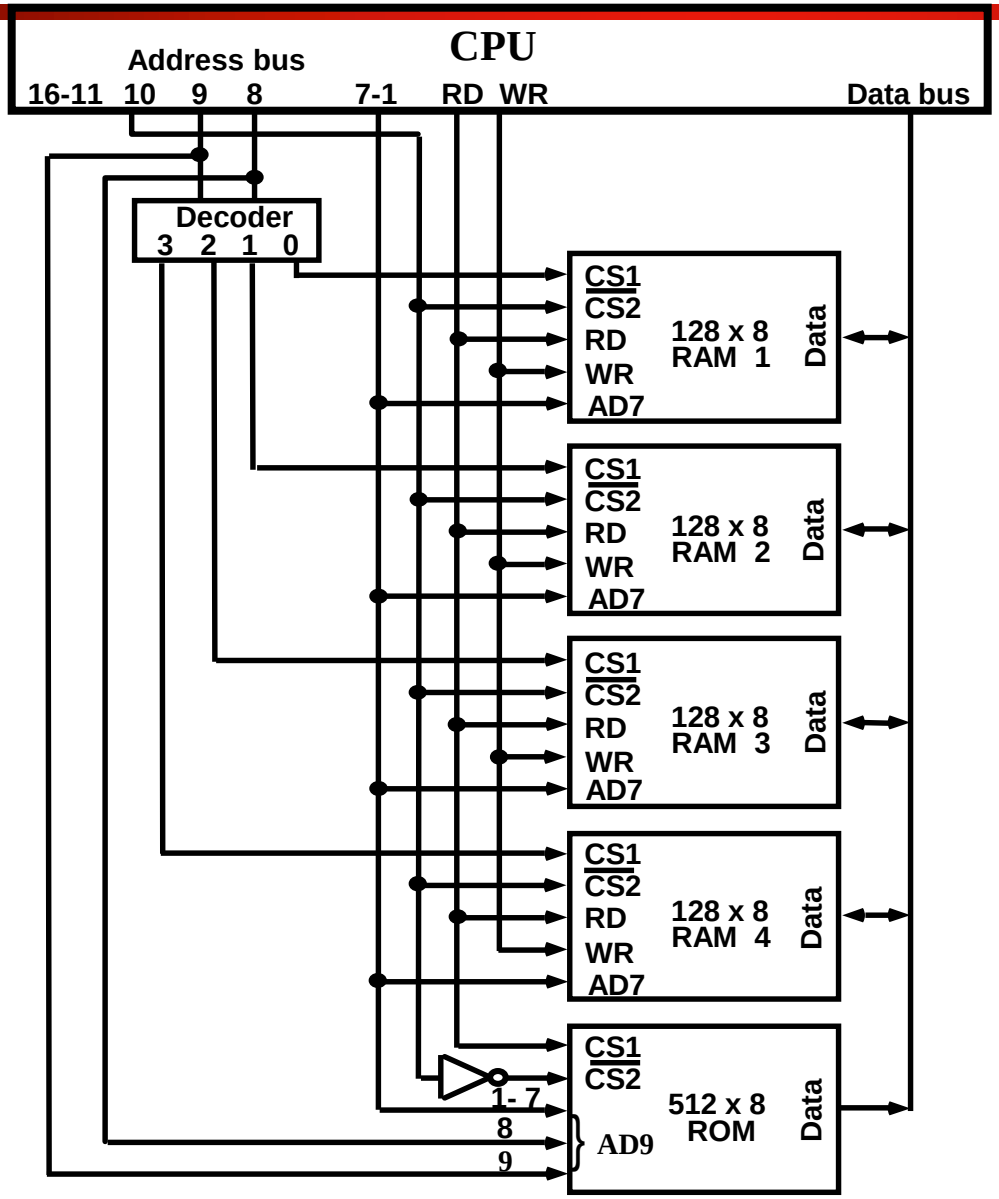
Component	Hexadecimal address	Address bus									
		10	9	8	7	6	5	4	3	2	1
RAM 1	0000 - 007F	0	0	0	x	x	x	x	x	x	x
RAM 2	0080 - 00FF	0	0	1	x	x	x	x	x	x	x
RAM 3	0100 - 017F	0	1	0	x	x	x	x	x	x	x
RAM 4	0180 - 01FF	0	1	1	x	x	x	x	x	x	x
ROM	0200 - 03FF	1	x	x	x	x	x	x	x	x	x

Memory Connection to CPU

- RAM and ROM chips are connected to a CPU through the data and address buses

- The low-order lines in the address bus select the byte within the chips and other lines in the address bus select a particular chip through its chip select inputs

CONNECTION OF MEMORY TO CPU



Numerical Problems

Q1. a) How many 128x8 RAM chips are needed to provide a memory capacity of 2048 bytes.

b) How any lines of the address bus must be used to access 2048 bytes of memory. How many of these lines will be common to all chips.

c) How many lines must be decoded for chip select? Specify the size of decoders.

Q2. Extend the memory system of Fig.1 to 4096 bytes of RAM and 4096 bytes of RAM. List the memory-address map and indicate what size decoders are needed.

Q3. A computer employs RAM chips of 256x8 and ROM chips of 1024x8. The computer system needs 2K bytes of RAM, 4K bytes of ROM, and four interface units each with four registers. A memory-mapped I/O configuration used. The two highest-order bits of the address bus are assigned 00 for RAM, 01 for ROM, and 10 for interface registers.

How many RAM and ROM chips are needed.

Draw a memory-address map for the system.

Give the address range in hexadecimal for RAM, ROM and interface.

Cache Memory

- 12-5 Cache Memory
  - Locality of Reference
    - the references to memory *tend to be confined within a few localized areas* in memory
  - Cache Memory : **a fast small memory**
    - keeping the most frequently accessed instructions and data in the fast cache memory
  - Hit Ratio
    - the ratio of the number of hits divided by the total CPU references (hits + misses) to memory
      - » ***hit*** : the CPU finds the word in the cache
      - » ***miss*** : the word is not found in cache (CPU must read main memory)
  - A computer with cache access time of 100ns, a main memory access time of 1000 ns, a hit ratio of 0.9 is having average access time of 200ns.

Types of Mapping of Cache Memory

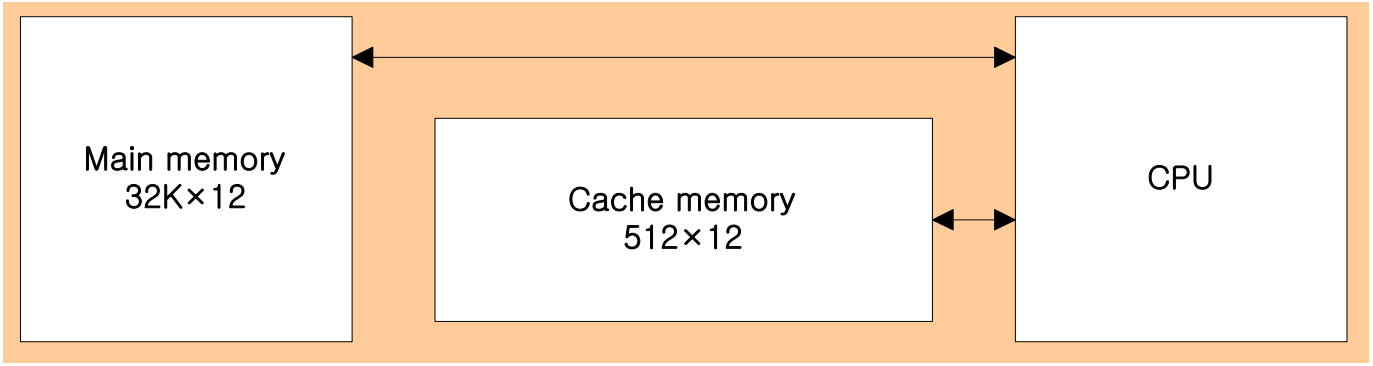
- Mapping
  - The transformation of data from main memory to cache memory
    - » 1) Associative mapping
    - » 2) Direct mapping
    - » 3) Set-associative mapping

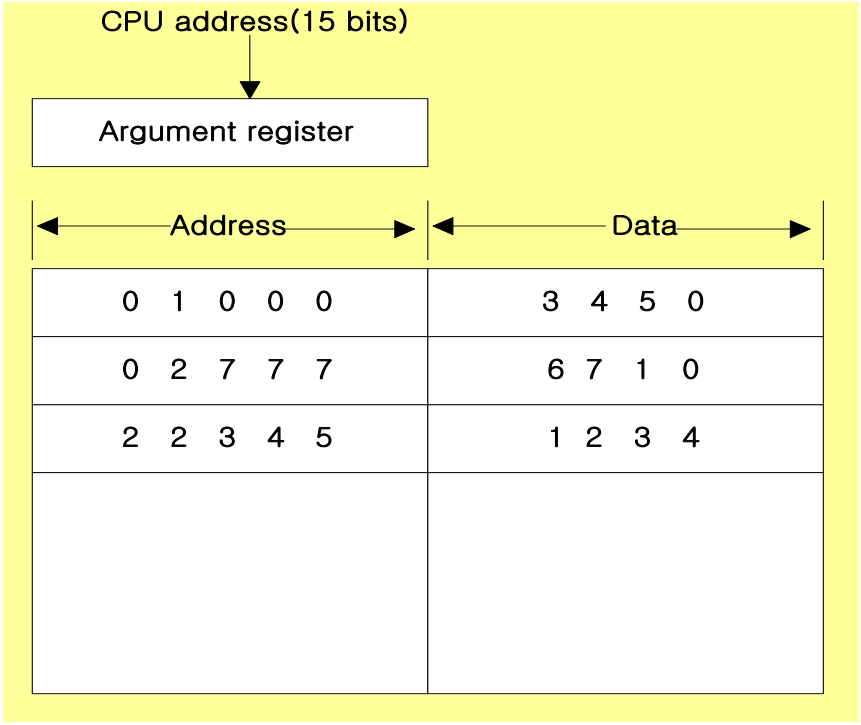
- Associative Memory:

  main memory : **32 K** x 12 bit word (15 bit address lines)

  cache memory : **512** x 12 bit word
    - » CPU sends a 15-bit address to cache
      - **Hit** : CPU accepts the 12-bit data from cache
      - **Miss** : CPU reads the data from main memory (then data is written to cache)

Associative Memory





Direct Mapping

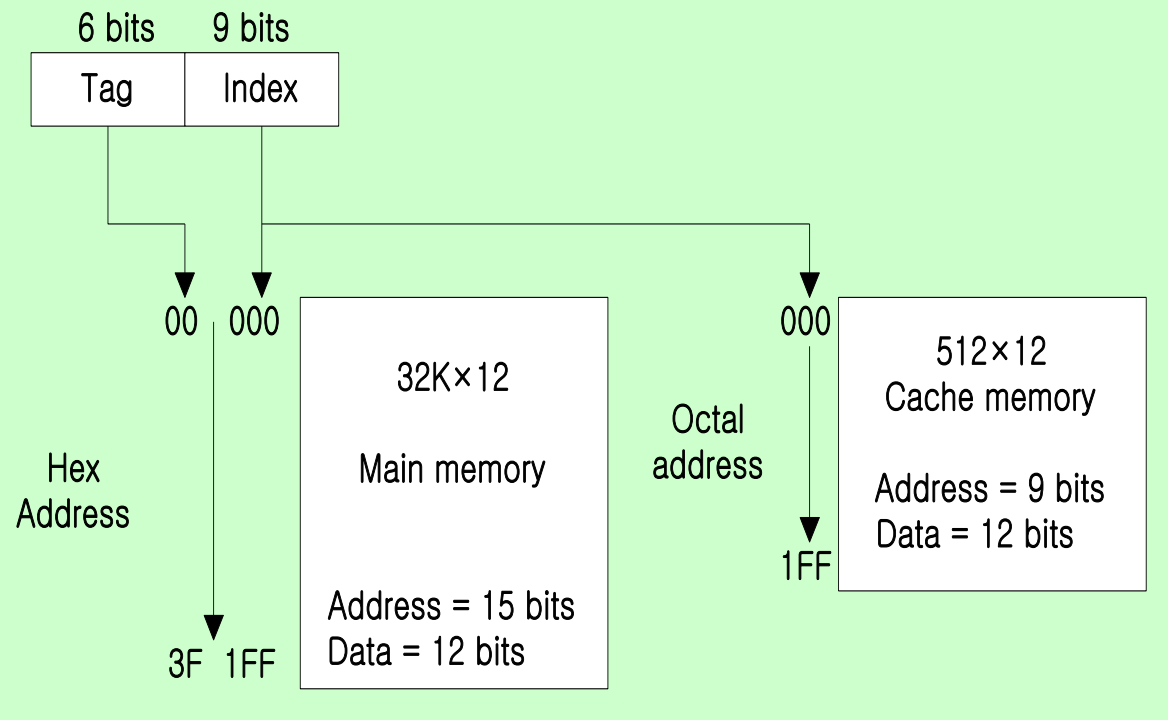
Direct mapping :

n bit memory address

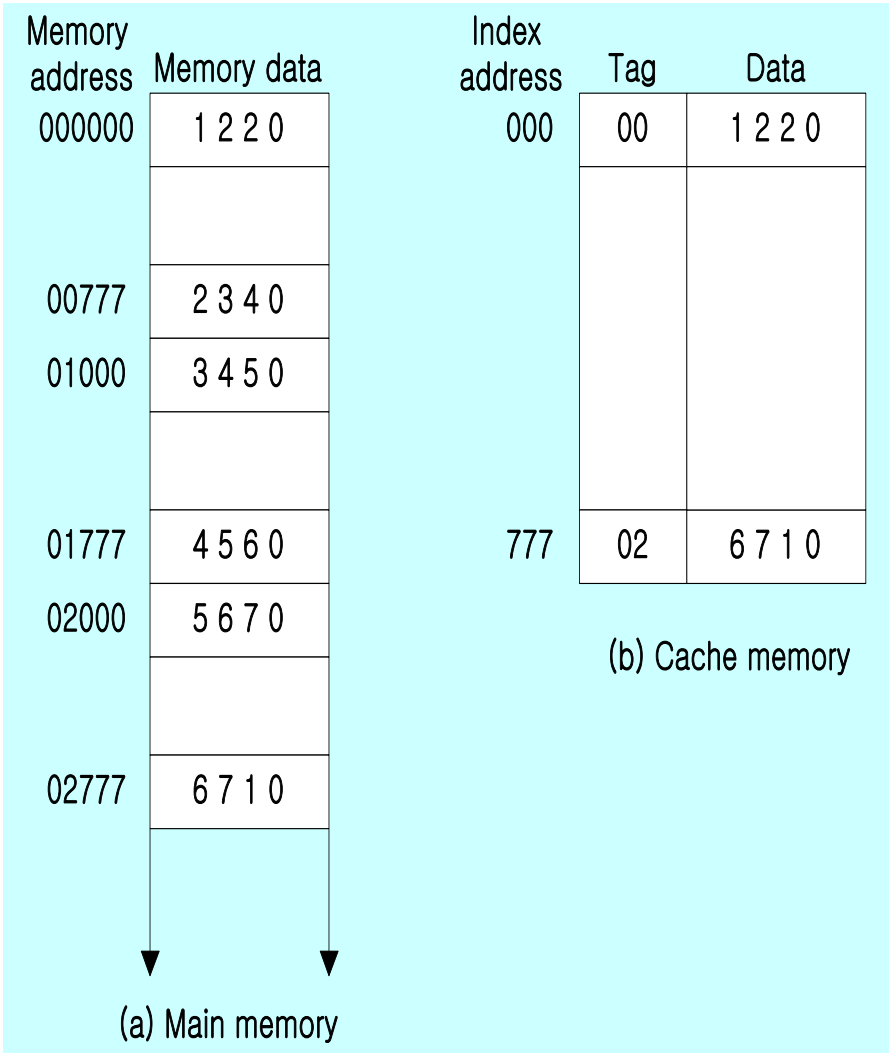
Tag field (**n - k**) : Index field (**k**)

2^k words cache memory + 2^n words main memory

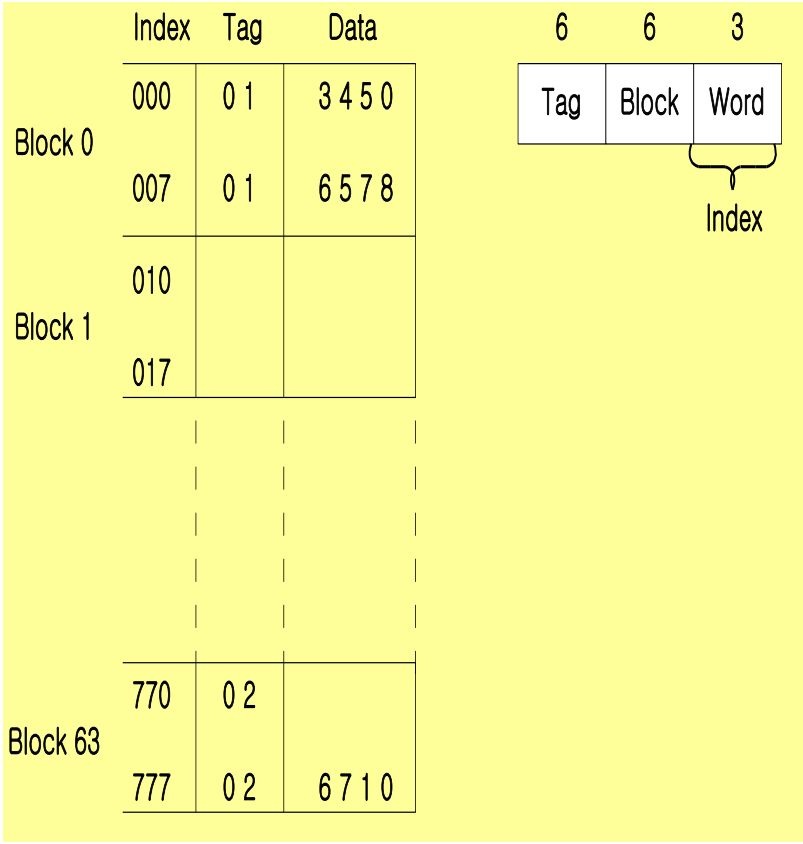
Tag = 6 bit (15 - 9), **Index** = 9 bit



Direct Mapping



- Direct mapping cache with block size of 8 words :
  - » 64 block x 8 word = 512 cache words size



Set-Associative Mapping

Set-associative mapping : (two-way)

Index	Tag	Data	Tag	Data
000	0 1	3 4 5 0	0 2	5 6 7 0
777	0 2	6 7 1 0	0 0	2 3 4 0

Questions

- A two-way set associative cache memory uses blocks of four words. The cache can accommodate a total of 2048 words from main memory. The main memory size is 128Kx32. Formulate all pertinent information required to construct the cache memory. What is the size of cache memory.

- A computer has a memory unit of 64Kx16 and a cache memory of 1K words. The cache uses direct mapping with a block size of four words. How many bits are there in the tag, index, block, and words fields of the address format. How many bits are there in each word of cache, and how are they divided into functions? Include a valid bit. How many blocks can the cache accommodate?

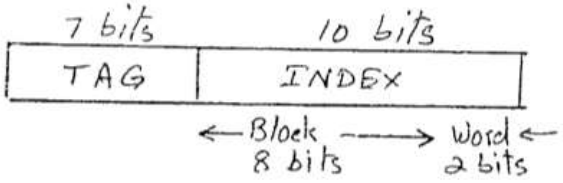
- The access time of a cache memory is 100ns and that of main memory is 1000ns. It is estimated that 80 percent of the memory requests are for read and the remaining 20 percent for write. The hit ratio for read access only is 0.9. A write-through procedure is used. What is the average access time of the system considering only memory read cycles. What is the average access time of the system for both read and write requests. What is the hit ratio considering the write requests also.

Answer

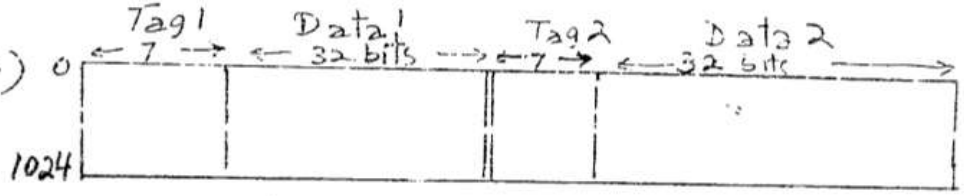
12-15

$128K = 2^{17}$; For a set size of 2, the index address has 10 bits to accomodate $\frac{2048}{2} = 1024$ words of cache.

(a)



(b)



Size of cache memory is $1024 \times 2(7+32)$
$= 1024 \times 78$

12-16

(a) $0.9 \times \underbrace{100}_{\text{cache access}} + 0.1 \times \underbrace{11000}_{\text{cache+memory access}} = 90 + 110 = 200$ nsec.

(b) $0.2 \times \underbrace{1000}_{\text{write access}} + 0.8 \times \underbrace{200}_{\text{read access from (a)}} = 200 + 160 = 360$ nsec.

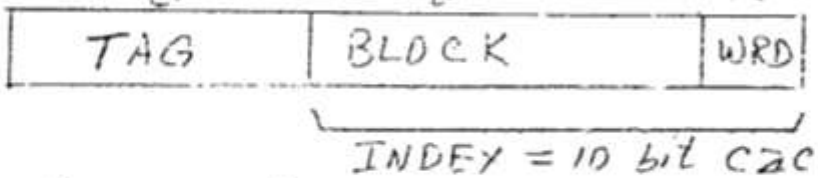
(c) Hit ratio $= 0.8 \times 0.9 = 0.72$

Answer

12-18

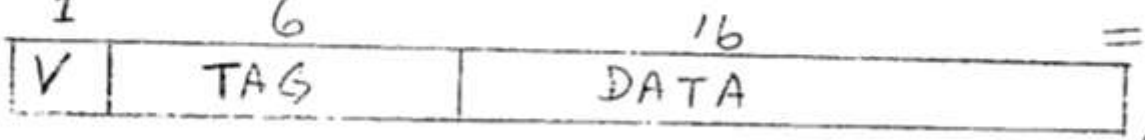
64K × 16 : 16 bit address ; 16-bit data.

(a)



= 16 bit address

(b)



= 23 bits in each word of cache

(c) 2^8 = 256 blocks of 4 words each

- Replacement Algorithm : cache miss or full
  - 1) **LRU** (Least Recently Used):
  - 2) **Random Replacement:**
  - 3) **FIFO** (First-In First-Out) :

- Writing to Cache :
  - » 1) **Write-through** :
  - » 2) **Write-back** :

- Cache Initialization
  - Cache is initialized :
    - » 1) When power is applied to the computer
    - » 2) When main memory is loaded with a complete set of programs from auxiliary memory
    - » 3) Cache is initialized by clearing all the valid bits to 0.
  - Valid bit
    - » Indicate whether or not the word contains valid data